Visualization of Tactile Features for Object Recognition with a Multi-Fingered Hand

S. Funabashi, G. Yang, F. Hongyi, A. Schmitz, L. Jamone, T. Ogata and S. Sugano

***Abstract*—Multi-fingered robot hands can be extremely effective to physically explore and recognize objects, especially if they are extensively covered with distributed tactile sensors. Convolutional Neural Networks (CNNs) have been proved successful in processing high dimensional tactile information and we introduced a Morphology-Specific CNN (MS-CNN) in which hierarchical convolutional layers which were formed following the physical configuration of the tactile sensors on the robot. However, why the network achieved high recognition rates of objects was not revealed. In this study, Grad-CAM++ as one of visualization methods which is suitable for CNN architectures is utilized. From the visualization result, we investigated which MS-CNN architecture enables the robot hand to successfully recognize 9 types of physical properties of objects by a single touch. A recognition rate of over 95% was achieved.**

I. Introduction

The state of the arts focus on CNNs for tactile sensors [1]. Even though CNNs have achieved prominent results, when it comes to multi-fingered hands, some hands have differently sized and shaped distributed sensors [2], which means that the question of how to input tactile information from such sensors to CNNs needs to be considered. In particular, the size and shape of the tactile patches on the hand varies, as well as the size of the fingers, which makes the implementation of CNNs difficult, as CNNs in general require rectangular-shaped inputs. It also makes difficult to extract multi-fingered level meaning from tactile information such as object grasping and in-hand manipulation due to its incapability of processing tactile information on multi-fingered hands at the same time. Previously, we introduced MS-CNN which processed the whole tactile information and Architecture III achieved the highest object recognition rate, and thus input maps should not be combined as architecture I but convolution layers should be combined unlike architecture IV[3]. However, how the network architecture works was not revealed, and thus what other tasks the network can be applied is not investigated, yet.

Firstly, we visualized the internal representations of the different network structures using Grad-CAM++ [4], which has been used for tactile sensors [5] but not for multi-fingered hands, yet. Why and how such representations support tactile tasks related to multi-fingered hands was investigated. Also, we demonstrated how these learned representations permitted efficient transfer learning from recognition of object instance to recognition of physical properties with complex contact states on several fingers.

Satoshi Funabashi os with the Waseda University, Future Robotics Organization, Waseda University, Okubo 3-4-1, Shinjuku, Tokyo 169-8555, Japan. (e-mail: s-funabashi@aoni.waseda.jp).

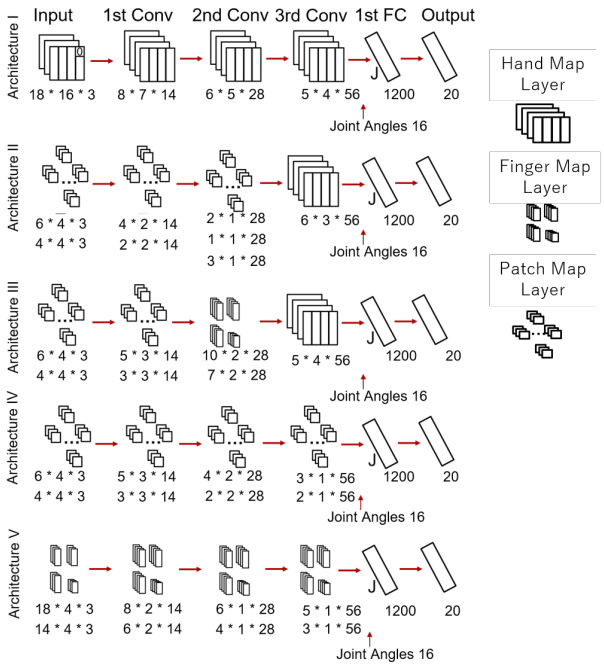

Fig. 1: Five architectures for combining convolution layers.

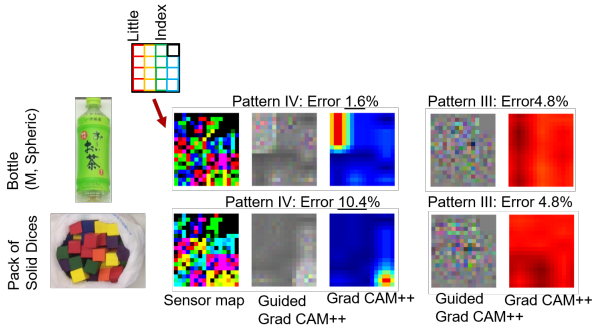

Fig. 2: Saliency maps from the last convolution layer in each MS-CNN generated by Guided Grad-CAM++ and Grad-CAM++.

II. Experiment Design Evaluation

In this study, we used the Allegro Hand, a commercially available robotic hand from Wonik Robotics and uSkin sensors. For the training data of object recognition, we used the same data as in our previous paper [3][6], where we confirmed the effect of uSkin sensor on multi-fingered hand by focusing on time series information and spatial information to improve the accuracy of object recognition by CNNs.

Fig. I shows network architectures used in this study. As

Architecture Pattern	Accuracy / Variance（%）	
	No transfer	With transfer
Architecture I	86.78 / 7.65	96.96 / 4.44
Architecture II	88.07 / 0.45	96.29 / 2.73
Architecture III	89.20 / 0.88	98.27 / 3.65
Architecture IV	82.87 / 2.35	96.07 / 5.12
Architecture V	81.80 / 2.90	93.60 / 7.18

Fig. 3: Recognition rate of the object property.

shown in Fig. 2, Bottle (M, spheric) and 12. Pack of solid dice were chosen for investigating how convolution layers affect recognition results. For the Pack of solid dices, architecture III got less error rate of object recognition than that of architecture IV. On the other hand, architecture IV got less error rate of object recognition than that of architecture III for the Bottle (M, spheric). The difference between the architectures is whether only the 'Patch Map Layer' is used or not. How the weights in the last convolution layer (3rd Conv layer) in each architecture react to tactile measurements was investigated. For architecture III, the saliency map provided by Grad-CAM++ shows tactile information is wholly focused. On the other hand, architecture IV shows the layer focuses on small part of tactile information, relatively. It seems that these differences happen because one or several convolution layers in the last layer (3rd Conv layer) were weighed heavily among convolution layers. For the grasped objects, the Pack of solid dices has relatively a complicated shape compared to the Bottle (M, spheric). This shape can change contact patterns on each part of sensor patch on the multi-fingered hand. For example, when the hand grasps the Pack of solid dices, contact patterns on each finger segment can be different which are provided by an edge, side and plane of the dices. On the other hand, the Bottle (M, spheric) has a small size enough that the hand grasps it wholly and a cylindrical shape which can produce a similar contact pattern on the hand while grasping. For this point, we deduce that it was easy for architecture IV to recognize relatively simple shaped objects. The network focuses on small part of contact areas and that can be enough to perform object recognition due to the similar contact pattern on any part of contact areas on the Allegro Hand. However, when it comes to complicated shaped objects such as the Pack of solid dices, they are difficult for architecture IV to recognize because it focuses on small contact areas, but contact patterns on the contact areas are diverse. For this result, we built the hypothesis that the CNN which has combined convolution layers (i.e. 'Hand Map Layer') focus entire tactile information on the multi-fingered hand and are robust to recognizing complicated contact states.

A. Object property recognition and transfer learning

For the result of the visualization, object property recognition during dynamic in-hand manipulation was targeted as it embraces complicated contact states. We deduced that the proposed CNN (especially architecture III) was robust to the contact states. The target in-hand manipulation was a movement of a precision grasp to a power grasp as starting with picking motion of fingertips from a ground.

There were 45 objects for object property recognition. The objects were separated into 3 heaviness classes (Heaviness High (more than 136g), Heaviness Medium (77 – 114g), Heaviness Low (under 68g)), and 3 softness classes (Softness High (deformable), Softness Medium (only surface deformable), Softness Low (stiff objects)), and 3 slipperiness classes (Slipperiness High (plastic or coated paper), Slipperiness Medium (paper or bumpy), Slipperiness Low (textile or rubber)) resulting in 27 classes.

The CNNs were trained with 13,094 samples which is smaller than 23,490 samples in the training dataset for the object property recognition task without transfer learning. As a result shown in Fig. 3, when networks (architecture I, II and III) had a 'Hand Map Layer', the networks got better result than the others and the architecture III got the best recognition result. Therefore, the proposed combined convolution is useful for the complex tactile information. Also, transfer learning was executed so that the CNNs could get better results even with smaller datasets. The architectures got around 10% better recognition rates and still the architecture III got the best result.

III. Conclusion

This study investigated how the MS-CNN architecture affected tactile-based multi-fingered hand tasks with distributed 3-axis tactile sensors by a visualized localization skill of the Grad-CAM++. Object recognition and object property recognition were targeted for the investigation. The visualization revealed that the proposed CNN is robust to complicated contact states on the robot hand. The best object property recognition rate was achieved by architecture III. Moreover, the proposed CNN achieved high object property recognition rates of 98% with transfer learning. Therefore, the convolution layers for each tactile sensor patch should be combined (not initially but gradually) to focus whole tactile information.

References

[1] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300–3307, 2018.

[2] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, "A review of tactile information: Perception and action through touch," *IEEE Transactions on Robotics*, vol. 36, no. 6, pp. 1619–1634, 2020.

[3] S. Funabashi, G. Yan, A. Geier, A. Schmitz, T. Ogata, and S. Sugano, "Morphology-specific convolutional neural networks for tactile object recognition with a multi-fingered hand," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 57–63.

[4] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 839–847.

[5] G. Cao, Y. Zhou, D. Bollegala, and S. Luo, "Spatio-temporal attention model for tactile texture recognition," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 9896–9902.

[6] S. Funabashi, S. Morikuni, A. Geier, A. Schmitz, S. Ogasa, T. P. Torno, S. Somlor, and S. Sugano, "Object recognition through active sensing using a multi-fingered robot hand with 3d tactile sensors," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 2589–2595.